Briefings in Bioinformatics, 00(0), 2021, 1–15

doi: 10.1093/bib/bbab185
Problem Solving Protocol

A multi-modal data harmonisation approach for discovery of COVID-19 drug targets

Tyrone Chen, Melcy Philip, Kim-Anh Lê Cao and Sonika Tyagi

Corresponding author: Sonika Tyagi, School of Biological Sciences, Monash University, 25 Rainforest Walk, 3800, VIC, Australia; Monash eResearch Centre, Monash University, 15 Innovation Walk, 3800, VIC, Australia; Department of Infectious Disease, Monash University, 85 Commercial Road, 3004, VIC, Australia; E-mail: sonika.tyagi@monash.edu

Abstract

Despite the volume of experiments performed and data available, the complex biology of coronavirus SARS-COV-2 is not yet fully understood. Existing molecular profiling studies have focused on analysing functional omics data of a single type, which captures changes in a small subset of the molecular perturbations caused by the virus. As the logical next step, results from multiple such omics analysis may be aggregated to comprehensively interpret the molecular mechanisms of SARS-CoV-2. An alternative approach is to integrate data simultaneously in a parallel fashion to highlight the inter-relationships of disease-driving biomolecules, in contrast to comparing processed information from each omics level separately. We demonstrate that valuable information may be masked by using the former fragmented views in analysis, and biomarkers resulting from such an approach cannot provide a systematic understanding of the disease aetiology. Hence, we present a generic, reproducible and flexible open-access data harmonisation framework that can be scaled out to future multi-omics analysis to study a phenotype in a holistic manner. The pipeline source code, detailed documentation and automated version as a R package are accessible. To demonstrate the effectiveness of our pipeline, we applied it to a drug screening task. We integrated multi-omics data to find the lowest level of statistical associations between data features in two case studies. Strongly correlated features within each of these two datasets were used for drug–target analysis, resulting in a list of 84 drug–target candidates. Further computational docking and toxicity analyses revealed seven high-confidence targets, amsacrine, bosutinib, ceritinib, crizotinib, nintedanib and sunitinib as potential starting points for drug therapy and development.

Key words: COVID-19; SARS-CoV-2; machine learning; multi-omics; data integration; data harmonisation; multivariate analysis.

Introduction

A new strain of coronavirus SARS-Cov-2 was identified in late 2019, which is responsible for the disease COVID-19 and the subsequent global pandemic. A significant amount of resources have been invested into studying this virus in order to limit its spread and negative impact on humans, resulting in a large quantity of publicly accessible epidemiological and molecular data (104 872 articles indexed in PubMed at this time of writing) [7]. While many of these molecular studies generate valuable data [3, 11, 19], we consider that a large quantity of existing studies observe data from a narrow perspective, generating a

Tyrone Chen is a PhD candidate in computational biology at the School of Biological Sciences, Monash University, Australia. His interest is developing and applying tools to uncover knowledge from biological data.
Melcy Philip is a bioinformatics postgraduate and research assistant. Her expertise is structural bioinformatics and molecular docking simulations.
Kim-Anh Lê Cao is an associate professor in statistical genomics at the School of Mathematics and Statistics, The University of Melbourne, Australia. Her lab focuses on the development of computational and integrative multivariate methods for feature selection in biological data.
Sonika Tyagi is a Machine Learning Lead and Senior Lecturer at the Monash University and the Alfred Hospital, Melbourne, Australia. Her research focuses on multi-modal data integration and machine learning applications in genomics and healthcare.
Submitted: 26 October 2020; **Received (in revised form):** 9 March 2021

single, specific category (or modality) of data corresponding to a single *omics* type.

Single-omics or 'unimodal' views of data contrast strongly with the known heterogeneity of biological systems. Complex traits and diseases such as COVID-19 are often a result of composite interplay between the genome, environment and multiple layers of functional genomics, for example the lipidome, metabolome, proteome and transcriptome. Highly complex signalling networks arise as a result of these interactions, and it is rarely straightforward to understand how their different components interact to produce a phenotype. High-throughput data generated from multiple functional layers of a biological system is known as 'multi-omics' or 'multi-modal' data that can be generated from the same of different cohorts of samples.

Accordingly, we consider the possibility of obtaining additional and novel information by integrating multiple omics datasets together. We define this as a 'multi-modal harmonisation' approach to homogenise and analyse data on the same scale, which is expected to capture a holistic view of the biological system under study, as opposed to more conventional sequential data aggregation or merging. Predicted advantages include greater data resolution, reduced noise and the ability to answer questions that a single data modality cannot, as demonstrated by existing studies [2, 18, 43]. Furthermore, the user will also have a higher degree of confidence in the results due to their concordance on separate data categories.

Data analysis is often performed on an individual, highly nuanced omics dataset using context-specific bioinformatics pipelines. Pipeline specificity, along with the significant differences across different omics data, hinders their direct comparison under normal circumstances. Generally, high-level data integration is performed after quantitative information across datasets have been reduced to a set of qualitative data, often resulting in a list of biological pathways. At this point, biological signal is weakened due to this information loss. Therefore, approaches that can unify and compare datasets simultaneously are favourable. In this article, we will be using the term 'harmonisation' [9] to refer to multi-modal data integration for finding the lowest level of statistical association between features of multiple data type.

We have previously reviewed and labelled data harmonisation strategies [9] that fall into two broad categories: (i) methods with restricted scopes impose specific assumptions and operate on a specific combination of omics data only and are of limited use in our data analysis context; (ii) methods with unrestricted scopes include less constraints (such as method-specific assumptions and data transformations) and can be subdivided into supervised and unsupervised methods. Supervised methods require the outcome, in this case, biological category, to be known while unsupervised methods such as JIVE [27], iCluster [38], MOFA [1], seurat [41], LIGER [48] NMF [54], iNMF [53] and SNF [46] do not. However, the greater flexibility of unsupervised methods is balanced by their lower classification performance relative to supervised methods [39]. Since the biological categories in our multi-omics dataset are known, we considered supervised methods. Among these methods, NetICS [12] and DeepMF [6] require prior information or manual parameter tuning. In comparison, Data Integration Analysis for Biomarker discovery using a Latent cOmponent (DIABLO) [39] does not have these disadvantages. An additional advantage of DIABLO is that it reports low-level feature associations across omics data.

At the same time, we identified a significant gap in the field. While a few methods exist to address the problem of low-level feature harmonisation, there does not yet exist an 'off-the-shelf' pipeline to perform this process. We filled this gap by writing an input-to-feature pipeline applying state of the art algorithms in data harmonisation [35, 39] and made it publicly available as a git repository and an R package [8].

We present two multi-omics case studies to illustrate one possible application of our pipeline on drug screening. We focused on harmonising low-level biological features across (i) a dataset with SARS-CoV-2 proteome and translatome data and (ii) another dataset with SARS-CoV-2 lipidome, metabolome, proteome and transcriptome data. Low-level biological features in this context correspond to individual biological molecule identities, for example a lipid, metabolite, peptide or transcript. In each of our case studies, we compared these features across each omics measurement directly, in contrast to most existing methods that are prone to information loss from simplifying data to a high level for comparison.

To show the relevance of our pipeline in biomedical applications, we applied it in the context of drug screening. As part of our downstream processes, we took correlated features output by our pipeline and applied prior knowledge as well as computational modelling as downstream analyses. Among existing drug databases, we selected DrugCentral [44] for its open accessibility and ease of use. For performing molecular docking simulations of drug–target combinations, we selected four computational methods, SWISSDOCK [14], PATCHDOCK [37], MTiAUTODOCK [22] and Achilles Blind Docking [36]. While each method uses a different strategy, our goal is to obtain a consensus among the docking methods to refine our final list of screened drugs. At the same time, we independently validated our drug associations with SARS-CoV-2 by exploiting the wealth of SARS-CoV-2 literature and clinical trials [Table S1]. We considered that concordance among all the above factors would result in a higher degree of confidence in the final results.

We supplemented these downstream processes with a literature survey of reported COVID-19 drug–target associations. We showed that many of these studies involved *in silico* validation of COVID-19 related drug–target combinations, implicating existing drugs or their analogues as potential drugs against SARS-CoV-2 infection [45]. Investigating these further revealed that clinical trials on many of these drugs are on-going [Table S1]. Most of our surveyed drugs target the SARS-CoV-2 main protease, where others targeted host proteins. Within this realm, few drugs were capable of acting on both the virus and host. Some of the popular methods for identifying the drug–target combinations are molecular docking, implementation of machine learning and deep learning techniques and pathway analysis.

In our pipeline, we apply a multivariate approach to harmonise multi-omics data and detect signals contributing to the viral state in two COVID-19 case studies. To illustrate the utility of our pipeline, we show that a list of features can be generated for downstream analysis from generic input data. To demonstrate its effectiveness, we show that it recapitulates published results from algorithms that were specially designed for these respective studies. From this, novel information on the molecular mechanisms of SARS-Cov-2, which is not possible to obtain with individual omics data, is applied to a drug screen. Furthermore, the generalisability of our pipeline makes it highly adaptable to investigating quantitative multi-omics datasets and is not restricted specifically to a SARS-CoV-2 system [10].

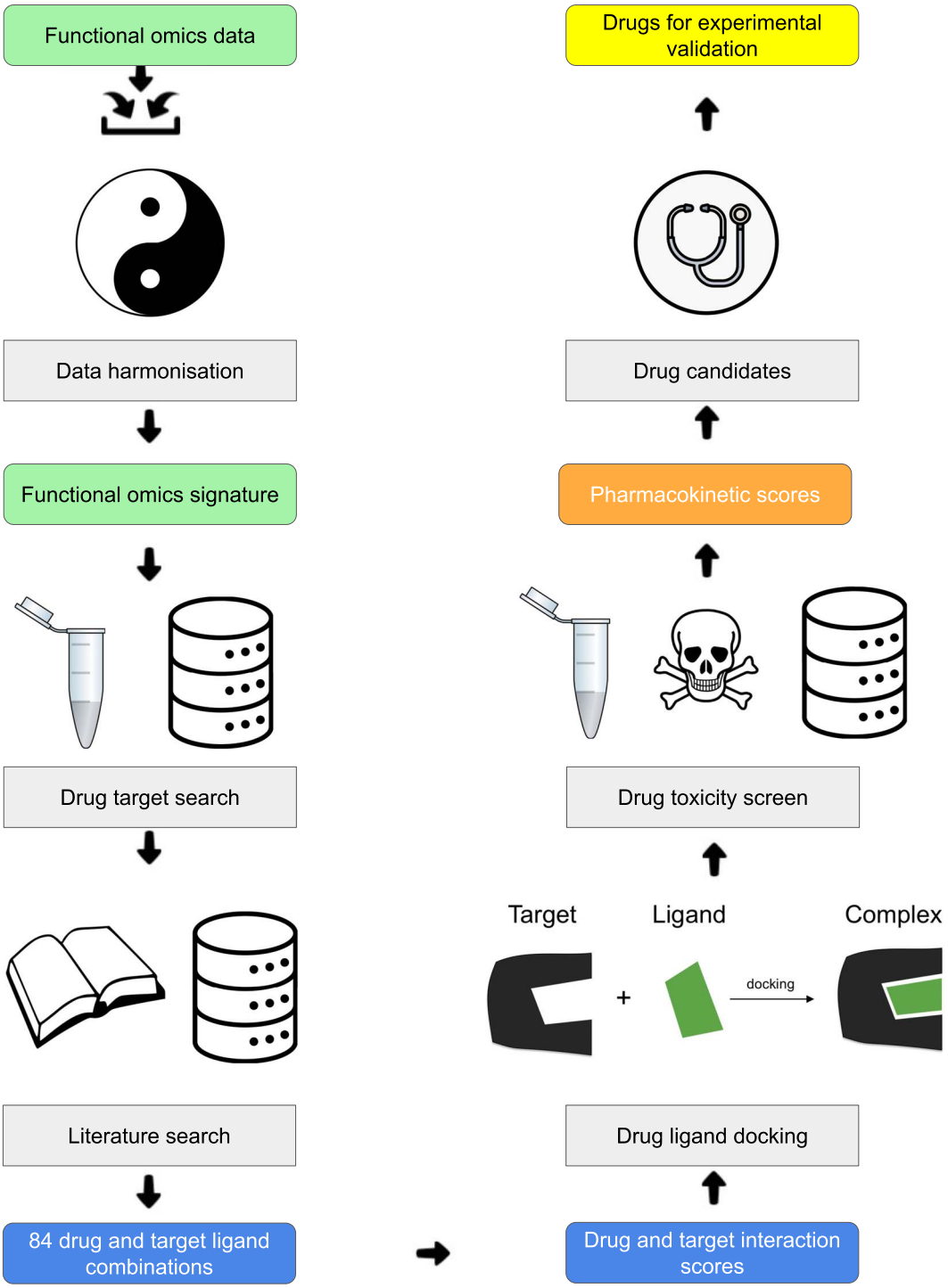

Figure 1. Graphical abstract of the publication.

Materials and methods

Data review, processing and analysis

Data selection

When selecting test case datasets for our pipeline, we considered several criteria. First, the data should contain at least two separate omics data. Second, the data would be quantitative. Third, the data would contain at least two biological classes or treatment conditions. Fourth, the data must be publicly accessible in a machine-readable format. Fifth, metadata must be similarly accessible and unambiguous. For the system under study, we narrowed our scope to datasets associated with SARS-CoV-2 to highlight the utility of our pipeline in a drug screening example. We surveyed datasets that met all criteria simultaneously within this system and selected two specific case studies. However, we note that the pipeline is generic and applicable across systems.

Table 1. Experimental design of case study 1 [3]. There are three biological replicates and eight treatment conditions, for a total of 24 samples in the whole dataset

Infection state	Timepoint (hour)
Control	2
Control	6
Control	10
Control	24
Infected	2
Infected	6
Infected	10
Infected	24

Table 2. Experimental design of case study 2 [31]. There are 49–51 biological replicates and 2 treatment conditions, for a total of 100 samples reanalysed

Infection state
Less severe
More severe

Data description

Two independent case studies are featured, each involving SARS-CoV-2 data. Both case studies, including links to source data, detailed documentation, R code and all steps to reproduce our analyses, are in a publicly accessible git repository https://gitlab.com/tyagilab/sars-cov-2. Our raw R data objects are included and can be loaded directly for the user to inspect low-level technical details and the underlying structure of the data. All figures in this manuscript that are directly generated by our pipeline can be reproduced by running the script https://gitlab.com/tyagilab/sars-cov-2/src/make_manuscript_figures.R.

Case study 1 For our first case study, a multi-omics dataset containing proteome and translatome data is available from [3]. SARS-Cov-2 virus was grown in cell culture of human colon carcinoma cell line CaCo-2, using published methods [3] and performed in three biological replicate cell cultures, then repeated for uninfected controls. The supernatant of infected cells was repeatedly sampled at four different timepoints and was repeated for uninfected cells as a control, yielding a total of eight treatment conditions, and a total of 24 samples. Prior to filtering, the raw data contained a total of 6381 proteomic features and 2715 translatomic features.

Case study 2 A second case study incorporating four omics datasets on patients infected with SARS-Cov-2 is also available to illustrate the utility of our pipeline [31]. A total of 100 patients with SARS-CoV-2 were categorised based into two classes (more severe and less severe) based on their condition using a composite metric developed by the original authors of the study. Before filtering, raw data contained a total of 3357 lipids, 150 metabolites, 517 peptides and 13 263 transcripts. In both case studies, data preprocessing steps and the flow of data through the pipeline are conceptually identical.

Data preprocessing

Data were downloaded from the original publications and reformatted into matrices of values using the python pandas software package [32, 49]. Data filtering was carried out by removing features that were not represented in every sample class. The non-linear iterative partial least squares (NIPALS) algorithm was used to impute remaining missing data values where applicable (proteome and translatome in case study 1, transcriptome in case study 2) [35]. In case study 1, principal component analysis (PCA) before and after multilevel decomposition for cell culture was used to assess the effect of the longitudinal study design [26].

Multivariate data analysis

Filtered, normalised abundance measurements of omics features were provided as input data and sample group information as input metadata. In addition, cell culture information was provided to account for the effect of repeated measurements on the same cell culture in case study 1.

In this study, we used a specific class of partial least squares (PLS) analysis methods and a sparse variant, which performs internal feature selection rather than operating on the full data matrix [24]. PLS can be applied to discover relationships between two matrices of data and unlike many methods is suited to the case of $n << p$, where the number of variables or features p outnumber the number of samples n. Therefore, it is particularly useful in biological data analysis, where for example, a single biological sample may be associated with thousands of genes of interest.

We considered two types of supervised PLS approaches: PLS discriminant analysis (PLSDA and sparse PLSDA) to perform supervised analysis on each single omics dataset [23] and multiblock sPLSDA, also known as DIABLO [39], to integrate the multi-omics datasets. Generalising sPLSDA to multiple blocks of omics data allows us to capture the correlations between features across different omics data.

In case study 1, a multilevel decomposition was used to remove individual variation caused by repeated measurements before variable selection and classification are performed simultaneously in the multivariate models. Features of interest contributing to each condition are output, along with their contribution weights. Performance of the model was evaluated by observing error rate and supplemented by area under the curve generated by cross-validation. It was also possible to survey the stability of selected variables.

Required input parameters include the number of components in these latent variable models, distance metrics and the optimal number of features to select. We tuned these parameters using 'leave-one-out' cross-validation.

The outputs of sPLSDA and DIABLO are lists of features per omics data, which strongly discriminate between biological categories in the experiment. Features are first linearly combined into components, which enable us to reduce data dimension, similar to PCA. Variable loadings correspond to coefficients assigned to each variable when calculating each component (their absolute value indicate the importance of each variable to define a component) and are used as a metric for measuring their contribution towards a biological outcome or state. In the case of DIABLO, the features that are highly correlated across the different omics data are expected to be informative.

The mixOmics R software package and its set of methods was used for multivariate data analysis [33, 35]. Detailed documentation with examples are available at the mixOmics website: www.mixOmics.org.

Computational validation of drug–target combinations

To place this information in a medical context, the most distinguishing features returned by our pipeline were investigated for

their potential as drug targets. For this purpose, we selected several protein–ligand docking methods, SWISSDOCK [14], PATCHDOCK [37], MTiAutoDock [22] and Achilles Blind Docking [36]. Instead of using a single protein–ligand docking method, we used all methods and compared their output. Should a potential drug target be identified by all methods, a higher degree of confidence would be obtained. Finally, these drugs were also cross-referenced against existing studies and clinical trials.

SWISSDOCK (based on the EADOCK ESS algorithm) measures the binding affinities of ligand–target interactions. It generates multiple clusters with respect to the binding sites and ligand–target interactions [14]. SWISSDOCK is calibrated for drug design and is effective for small and rigid ligands. As input, Protein Data Bank (PDB) [4] identifiers of the drug and target names are required. Within input parameters, a flexibility of Åwas specified and no region of interest was selected. The server retrieves protein and ligand details from their respective databases to perform docking calculations, which results in the generation of the docked structure, a protein–ligand complex.

PATCHDOCK is a freely accessible web server and measures geometric complementarity for performing structure prediction of protein–small molecule complexes and protein–protein complex [37]. As input, '.pdb' files of the drug and targets are required. Within input parameters, clustering root mean square deviation was set to four (default). A series of molecular transformations were performed and the best shape complementarity was obtained.

MTiOpenScreen is a web server exclusively for molecular docking and virtual screening [22] of drug molecules by their lowest energy score. Docking is performed using MTiAutoDock and screening with AutoDock Vina [22]. As input, '.mol2' files of the drug and target are provided. Within input parameters, hetero atoms are replaced with hydrogen atoms by MGLTools [13].

The Achilles Blind Docking Server is available at http://bio-hpc.eu/software/blind-docking-server/ and implements the BINDSURF algorithm [36] to calculate binding energies. It performs a comprehensive iteration of docking calculations over the protein surface in order to find the pockets with best binding affinities [36]. Many different docking simulations are performed on each alpha carbon of the protein, enabling the detection of new binding spots. Results are clustered according to spatial overlapping of the listed drug–target poses and the highest affinity selected. As input, '.pdbqt', '.mol2' or '.pdb' files are accepted. Manual preparation was needed depending on the molecular attributes of the input.

In silico pharmacokinetic analysis of the drugs was done by analysing their absorption, distribution, metabolism, and excretion (ADME) properties and was carried out using admetSAR tool [52]. As input, simplified molecular input line entry system formatted entries of the drugs were submitted, and the results showed the viability of the drug. The tool requires no input parameters.

Crystal structure generation

A known crystal structure is required for docking analysis. A protein we identified (Uniprot ID: O94956) lacked a known crystal structure entry in the PDB database. Another protein (Uniprot ID: P35869) was too large for molecular docking analysis. Hence a crystal structure was generated using homology modelling for protein O94956, and the crystal structure for protein P35869 was modelled using only chain A of the same [47]. The templates were ranked based on the similarity, and the top one was used for the modelling of the proteins. The modelling was done using SWISSMODEL, which works based on target-template alignment using ProMod3 [47]. The output comprises the project report and final PDB structure of the modelled protein.

Software availability

All steps, code, parameters, command line arguments and software versions used to generate the multi-omics data integration results in this paper are publicly available in an open source software repository (MIT License) hosted on gitlab at https://gitlab.com/tyagilab/sars-cov-2. Documentation is available at the same location, licensed as CC-BY-3.0 AU. The snapshot of the repository that produced the data generated in this paper is available at https://doi.org/10.5281/zenodo.4562010.

All third-party software (SWISSDOCK [14], PATCHDOCK [37], MTiAutoDock [22], Achilles Blind Docking [36], admetSAR [52], SWISSMODEL [47]) used to validate drug–target combinations are published and available at their respective websites.

Data availability

For the first case study, translatome data are available from the source publication [3] as SupplementaryTable1 and proteome data are available as SupplementaryTable2. For case study 2, the authors provided their data in ansqldatabase.

For the second case study, all data are available as a set of sql tables [31]. Preprocessing code and documentation to reformat these data into a format compatible with our pipeline are available in our gitlab repository.

Results

Validating data quality before analysis

Infected cells express the ACE-2 receptor for SARS-CoV-2 entry

SARS-Cov-2 is known to use the angiotensin converting enzyme 2 (ACE-2) receptor to enter human cells [17]. The Caco-212 cell line selected in case study 1 expresses ACE-2 and is known to support growth of a related virus SARS-Cov-1 [30]. As an additional layer of validation, we examined the proteomics data in case study 1 and observed that ACE2 protein was present in all samples (not shown). Viral transcript load was also observed to increase over time (not shown), matching the authors' observations [3].

Filtering and imputing missing values

Before analysis, we assessed data quality to ensure that it was suitable for testing our hypothesis. We discovered a high proportion of missing values within the translatome data as well as a significant quantity of variation between individual cell cultures in all data. We addressed these issues with the following steps.

To overcome the missing value problem, we used a two-step approach. We first filtered out all features that were not represented at least once per class for a sample. As a result, in case study 1, the proportion of missing values was reduced from 47% to 17% within translatome data and did not affect proteome data. In case study 2, the proportion of missing values was 0.5% for transcriptomics data. Next, we then imputed these data with the NIPALS algorithm [42, 50, 51]. For consistency within case study 1, we repeated this with proteome data even though the proportion of missing values was low ($<$0.01%).

We next decided to investigate if the imputation introduced unwanted technical variation into the dataset, which would bias

any downstream analyses. To assess similarity, we used two subjective metrics, with the expectation that similarity between the unimputed and imputed data should be high. First, we examined the data with PCAs before and after imputation and noted no significant differences (Figure S1). Next, we plotted a heatmap of correlations between the principal components before and after imputation (Figure S1). The correlations between components were consistent before and after imputation for proteomics data, which is unsurprising given the small quantity of values imputed. Within translatomics and transcriptomics data, correlations were consistent across at least the top five components accounting for the majority of variation in the data.

Removing unwanted technical variation from primary data that might impact downstream analysis

We then considered the implications of case study 1 being longitudinal. Classical methods assume sample independence across the datasets and are therefore less appropriate in our context of analysis. To determine the extent of the effect of these repeated measurements on the same cell cultures, we first perform a multilevel decomposition as the first step of our analysis [26]. Using this to account for variation within the samples, we then performed a PCA of the decomposed data. For comparison purposes, we performed a PCA on the unadjusted data. As expected, we identified a strong cell culture effect in the original data, which was reduced in the decomposed data (Figure S2). Thus, having shown that the individual variation was present in the data, we adjusted for the repeated sample measurements in the remainder of the analyses.

Multivariate single-omics analysis of primary data reveals pathways associated with viral infection and cellular stress

We first used a single-omics strategy to identify features highly contributing to each biological condition surveyed. We applied PLSDA and its sparse variant sPLSDA on each individual omics dataset, while accounting for the repeated measurements within each cell culture in case study 1. Visualising the first few components of these methods reveals several distinct sample groupings, with the main distinction in case study 1 being samples 24 hours post-infection versus all other samples, while in case study 2 a spectrum of values is shown by disease severity (Figures 2, S3, S4, 3, S5 and S6). Across PLSDA and sPLSDA within both case studies, this pattern is similar, matching the independent observations of the laboratories that originally generated the data [3, 31]. For case study 1, multiple secondary distinctions are also visible between data groups in both omics datasets, mostly between groups of time points from either infected or uninfected samples. The resulting feature lists are also generated for downstream analyses such as pathway enrichment and are available in the git repository.

Multivariate multi-omics analysis of primary data reveals a harmonised multi-modal signature

Having assessed the major sources of variation and features of interest contributing to biological conditions within the individual blocks of omics data (Figures 2 and 3), we used this information to guide our multi-omics data harmonisation. In each case study, we applied DIABLO to identify a highly correlated multi-omics signature in the data [39]. The DIABLO analysis was carried out in a conceptually similar way to the previous sPLSDA, except with multi-omics data as input instead of single-omics. We illustrated the correlation between features across these omics blocks with circos plots (Figures 4 and S7).

Multi-omics analyses achieve a lower error rate than single-omics

To objectively assess the performance of PLSDA, sPLSDA and multi-block sPLSDA (DIABLO) in case study 1, we compare the error rate per component and show that in all cases error rate was low (Figures 5, S19, S11 and S12). Using 'leave-one-out' cross-validation, we showed in all single-omics analyses that four components in case study 1 were sufficient to obtain a classification error rate approaching 0 with the centroids distance metric. In the case of the combined datasets, we achieved a slightly better performance with eight components and the Mahalanobis distance metric. We repeated these steps for case study 2 and show that an analogous pattern is visible, where error rates for single-omics analyses are higher compared to multi-omics analyses with two components with the centroids distance metric.

As a supplementary layer of validation, ROC curves showing classification accuracy (Figures S13 and S14) are available, but we note that these have limited applicability in the context of the method, which already internally specifies the prediction cutoff to achieve maximal sensitivity and specificity [35].

Our generic pipeline recapitulates results from published studies using custom pipelines

We compared our highly correlating multi-omics feature values to those in our case studies and showed that our strongly correlated multi-omics features overlap with highly scoring features identified in the original analyses (Figure 6).

In the first case study [3], the authors provided features ranked by two-sided, unpaired t-tests with equal variance assumed. First, we filtered this list by removing all features that were below or equal to a *P*-value threshold of 0.05. Next, we matched the 170 features we discovered through our pipeline to these features implicated as significantly differentially abundant in the original analysis. We find that the median of our subset of 170 features fall below the median *P*-value scores identified in the original analysis, suggesting that our pipeline identifies a subset of the more significant features. In particular, our pipeline recovers SRSF10, MAVS and GSTP1, which are important proteins associated with key pathways highlighted by the original study [3] due to their roles in pre-mRNA processing pathways essential for SARS-CoV-2 replication, viral processes and apoptotic regulation, respectively.

In the second case study [31], the authors ranked features by an importance score with their tree-based classifier. First, we filtered this list by removing all features that were assigned an importance score of 0. Next, we matched the features we discovered through our pipeline to these features found by the original analysis. We find that a subset of 24 features fall above the median score of importance among important features identified in the original analysis. In particular, our pipeline recovers quinolinic acid, which was an important metabolite highlighted in the original study [31] due to its role in immune function [16, 40] and COVID-19 severity [29].

We note that in each case, our list of features is comparatively small compared to the full range of significant features identified by the original studies. However, we emphasise that our approach works by selecting subsets of features. When

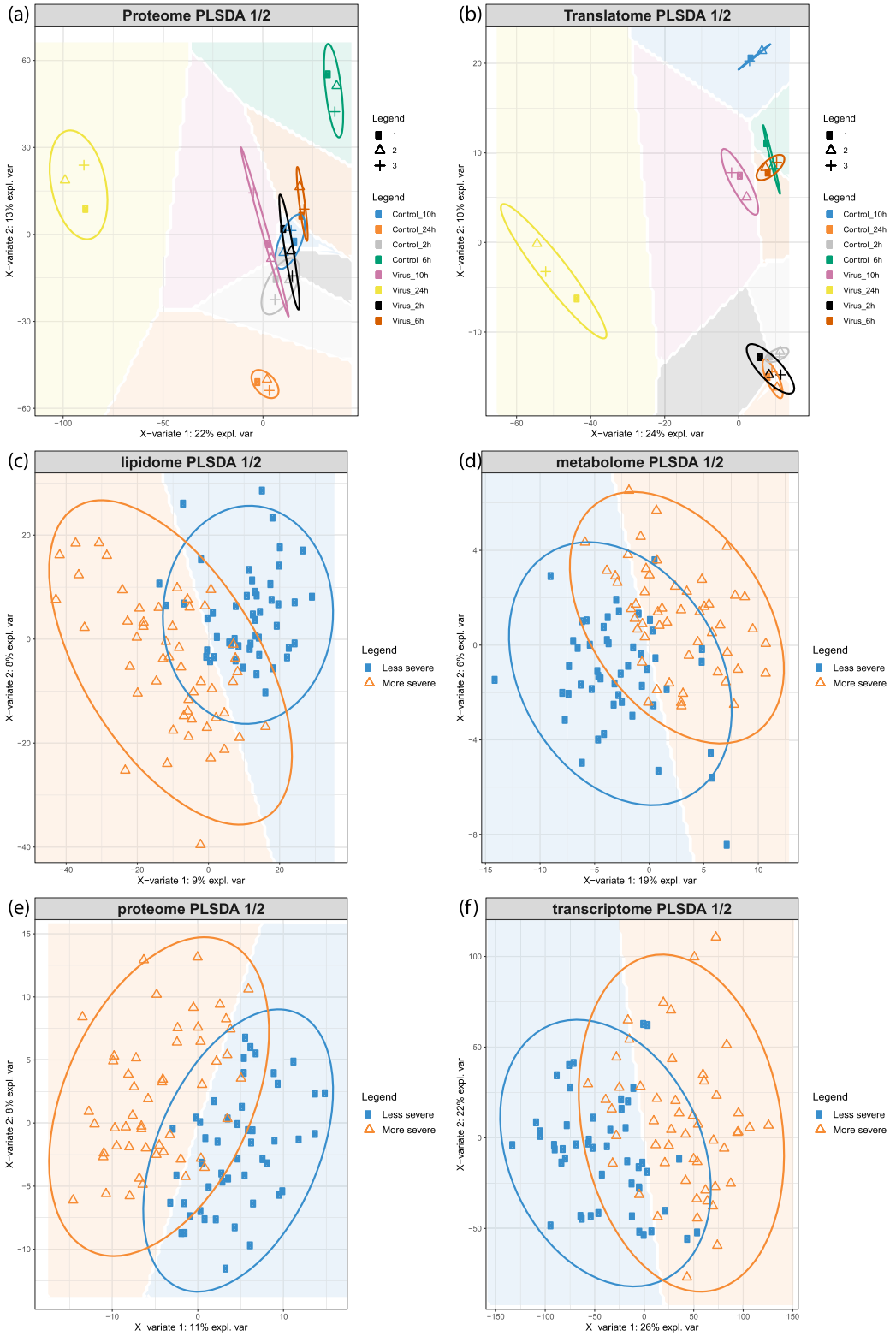

Figure 2. PLSDA plots on individual blocks of omics data. Case study 1: (A) proteome, (B) translatome; case study 2: (C) lipidome, (D) metabolome, (E) proteome, (F) translatome. The first component is plotted as the horizontal axis and the second component is plotted as the vertical axis, together accounting for approximately 20–50 % of the variation in the data across all cases. Background colour indicates prediction area. For case study 1, the main source of variation within each block of data appears to be the differences between samples 24-hour post infection and all other samples. In case study 2, a spectrum of cases on disease severity is visible.

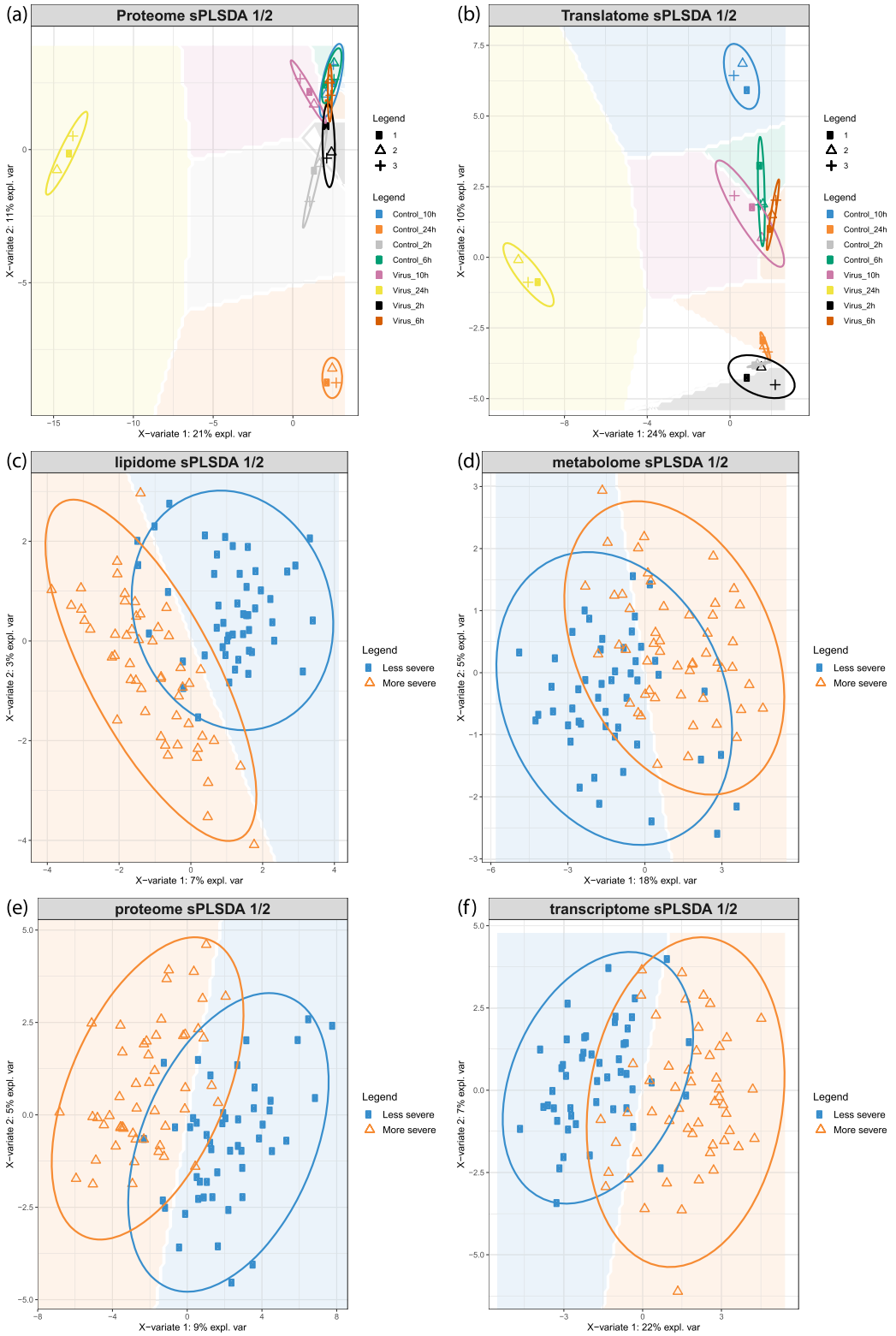

Figure 3. PLSDA plots on individual blocks of omics data. Case study 1: (A) proteome, (B) translatome; case study 2: (C) lipidome, (D) metabolome, (E) proteome, (F) translatome. The first component is plotted as the horizontal axis and the second component is plotted as the vertical axis, together accounting for approximately 20–50 % of the variation in the data across all cases. Background colour indicates prediction area. For case study 1, the main source of variation within each block of data appears to be the differences between samples 24-hour post infection and all other samples. In case study 2, a spectrum of cases on disease severity is visible.

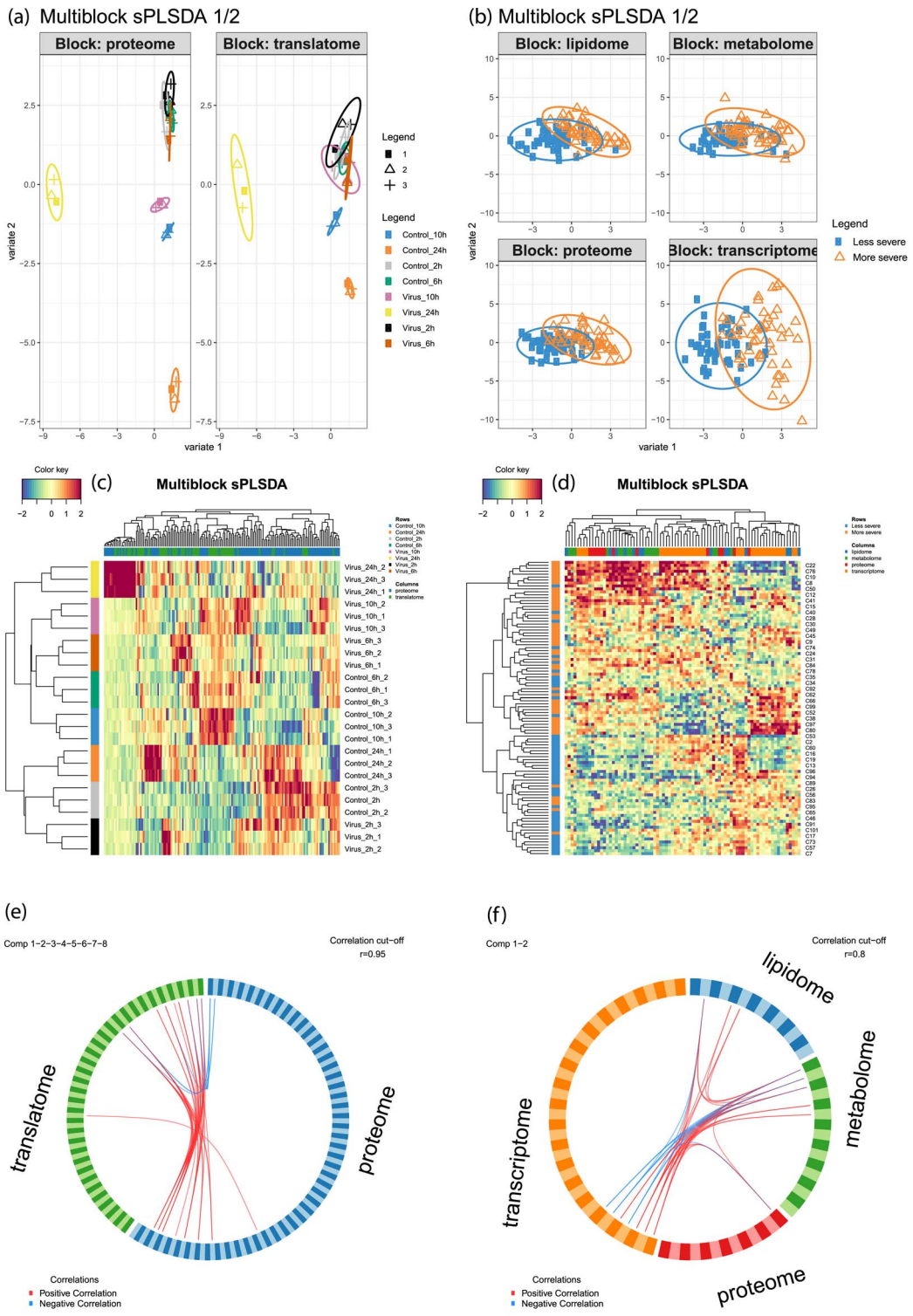

Figure 4. DIABLO (Multi-block sPLSDA) integrating multi-omics data. (A) Each block of omics data within case studies 1 (A) and 2 (B) is plotted for side-by-side comparison. A clustered image map shows the relationships across sample groups within case studies 1 (C) and 2 (D). Circos plot, built on a similarity matrix demonstrates the correlation between different proteins and transcripts, with a visualisation cutoff of 0.95 correlation score. Positive correlations are in red and negative correlations are in blue. For (E) case study 1, the proteome block is in blue and translatome block is in green. For (F) case study 2, the lipidome block is in blue, metabolome block is in green, proteome block is in red and transcriptome block is in orange. Line graphs on the outside of the circos plot represent expression levels of their corresponding features and are coloured by their biological sample class.

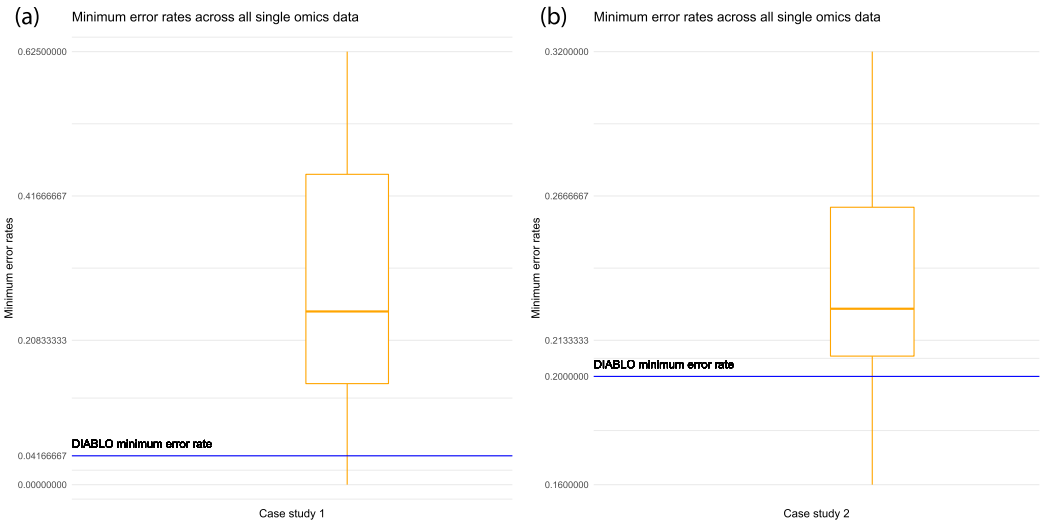

Figure 5. Minimum error rates obtained for each single-omics and analysis. Orange boxplots show the error rate distribution for single-omics analyses and overlaid blue line shows the minimum error rate obtained from the multi-omics analysis for (A) case study 1 and (B) case study 2. Note that further details on the individual error rates per omics data block are in Supplementary Figures S10, S11 and S12.

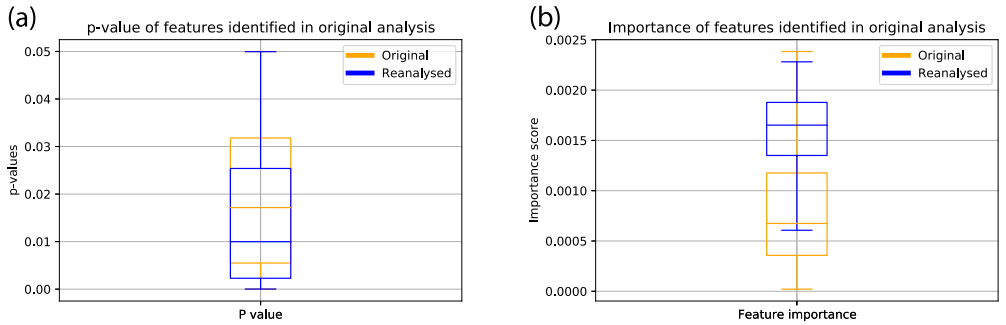

Figure 6. **Important features in original analyses recapitulated by our pipeline.** Orange boxplots show p-values (case study 1) and feature importance score (case study 2) distributions for the original analyses. Overlaid blue boxplots show the subset of features identified by our pipeline for (a) case study 1 and (b) case study 2.

contrasted against the original data, we illustrate that a subset of the most informative features were selected in each case. Furthermore, our generic pipeline was shown to partially recapitulate results obtained from two conceptually unrelated methods on separate experimental systems, including an algorithm that was specifically tailored to case study 2. Therefore, we demonstrate that our aim of providing an agnostic, flexible pipeline for analysing multi-omics data is achieved.

Predicting drug targets from a harmonised multi-modal signature

Having demonstrated that our multi-omics pipeline is effective, we exploit our findings from the reanalysis of both case studies to explore potential drugs for SARS-CoV-2. We first filtered our list of correlated multi-omic features on a threshold of ¡ -0.5 or ¿ +0.5 (Figure 4) and checked these against the DrugCentral database [44], resulting in the identification of 84 drug and target combinations [Table S1].

A literature survey reveals COVID-19 associations with our list of predicted drugs

Many of these drugs are already reported by existing preprints or publications. A large number of the drugs we reported are also undergoing clinical trials [Table S1]. We also report a few drugs that have not yet been identified by the literature, which may be of interest for further investigation. Among these are aspartic acid, asulacrine, carubicin and daunorubicinol. A possible explanation may be a limitation of the DrugCentral database [44], which records specific protein targets but does not at time of accession contain coronavirus entries.

Computational validation predicts the most likely drug–target interactions

To further validate our results and to characterise these proteins further, computational ligand–target docking analyses were carried out. Four independent methods SWISSDOCK [14], PATCHDOCK [37], MTiOpenScreen [22] and Achilles Blind Docking [36] were used to increase confidence in the results, as each algorithm approaches docking with a different strategy (Figure 7). Furthermore, we considered that a consensus among all algorithms would provide an additional layer of validity.

SWISSDOCK outputs the full fitness score and binding affinities of the ligand–target combinations. Lower fitness scores and binding affinities indicate stronger ligand–target interactions. For SWISSDOCK, etoposide phosphate, asulacrine and amsacrine had the highest scoring fitness scores in case study 1. Pazopanib, crizotinib and tranilast scored highly in case study 2. With the exception of tranilast, all of these drugs inhibit DNA, RNA and protein synthesis. Meanwhile

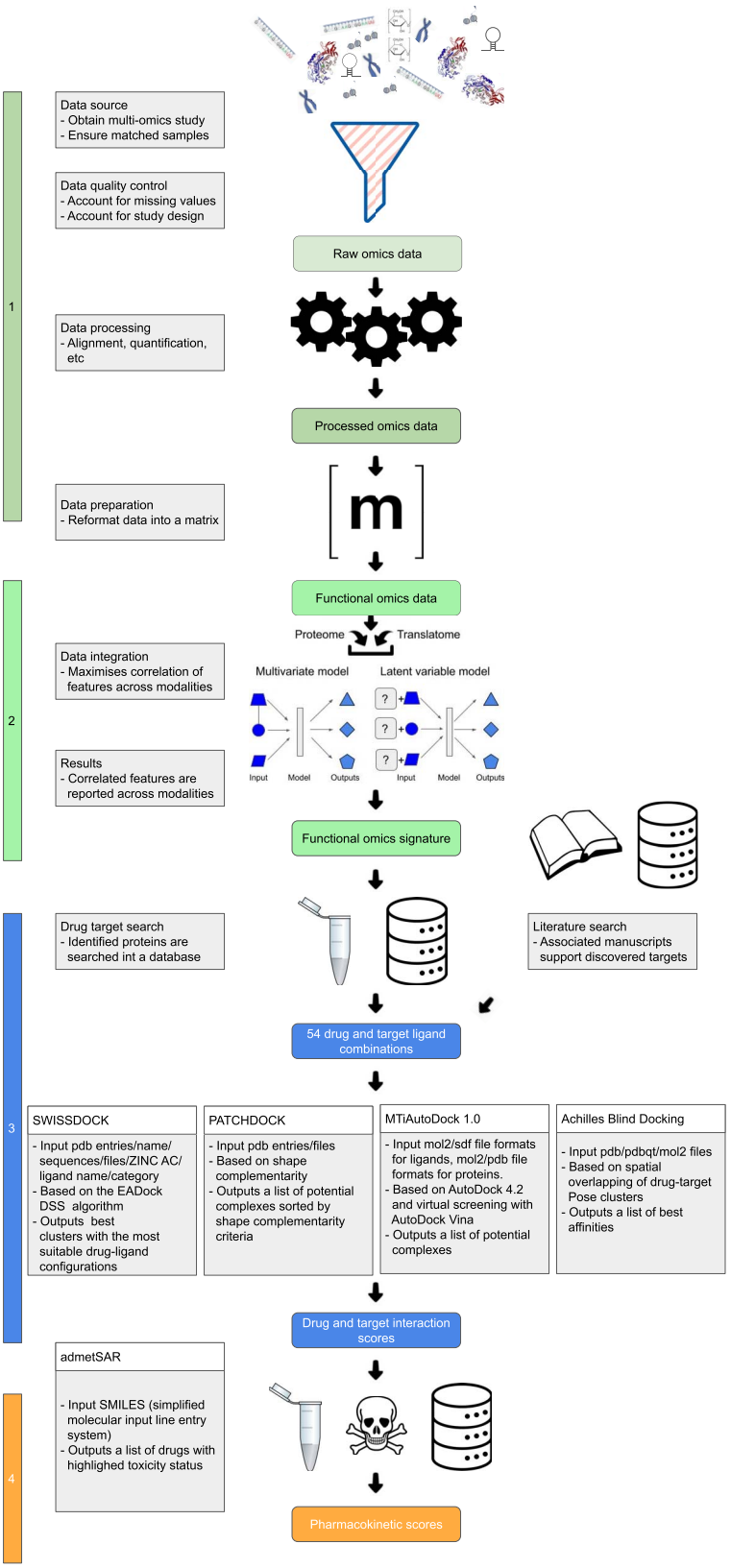

Figure 7. A flowchart visualising the steps taken in the pipeline. Functional omics data were obtained and filtered before harmonisation. The resulting protein list was searched against a drug database and a list of targets was identified. Using this list, docking analyses were performed with drug–target ligand combinations. These were further assessed for toxicity. Stage 1 demonstrates data selection as well as preparation, Stage 2 shows the main body of the pipeline and Stages 3 and 4 represent downstream analyses

PATCHDOCK reports ligand–target affinity in terms of geometric complementarity. The top few ligands reported by PATCHDOCK in case study 1 are erlotinib, ruxolitinib and afatinib, all of which are tyrosine kinase inhibitors. In case study 2, somatostatin, pasireotide and lanreotide are returned, all of which are growth hormone inhibitors. MTiAutoDock on the other hand searches for the ligand–target combination with the lowest complex energy score. Nintedanib, amsacrine and dasatinib are the best performing drugs in case study 1 by this criteria, while bosutinib, fentiazac and pasireotide are identified as highly scoring drugs in case study 2, all of which are growth hormone inhibitors, tyrosine kinase inhibitors or biomolecular synthesis inhibitors. Finally, Achilles Blind Docking returns results based on binding energy, where a lower score represents better ligand–target interactions. Achilles Blind Docking returns nintedanib, midostaurin and idarubicin as its best targets for case study 1 and sulfobromophthalein, nintedanib and midostaurin for case study 2. Compared to the top results of other algorithms, these drugs are more varied in their mechanisms of action, where midostaurin is a protein kinase inhibitor, idarubicin is a topoisomerase inhibitor and sulfobromophthalein is an enzyme inhibitor but not typically used as a drug.

Due to the diversity of objective metrics used by these molecular simulations, their results are not directly comparable. Therefore, we ranked their results and reviewed overlapping drugs across all methods. The full list and objective metrics are available [Table S1]. Three drugs, amsacrine, ceritinib and crizotinib, are common across all top ranked lists in case study 1. In case study 2, bosutinib, nintedanib, midostaurin and sunitinib are common. Examining these further shows that amsacrine disrupts DNA topoisomerase II activity, while bosutinib, ceritinib, crizotinib, nintedanib, midostaurin and sunitinib target and inhibit various kinases, subsequently suppressing cell growth and division. In addition, these drugs are also used to suppress tumour growth in cancer. We investigated existing work to gain insight into this observation and found an independent study on similar biological material, which discovered that infection triggers cell growth pathways, and drugs suppressing these inhibit viral replication [20].

Regarding the remaining candidate drugs, a careful review of existing studies that predict COVID-19 related drug associations supplements these findings. Each of these drugs have a reported association with SARS-CoV-2 in either a peer-reviewed study, preprint or clinical trial. Most of the drugs were found to target the main protease of SARS-CoV-2, and some target host proteins. Few drugs acted on both virus and host targets. We observed that many of these drugs are already in clinical trials for COVID-19.

Most drugs are predicted to have physiochemical properties suitable for humans

To obtain additional evidence supporting drug suitability with human physiology, we examined drug pharmacokinetic properties. Adsorption, distribution, metabolism, excretion and toxicity (ADME/T) properties of the drug were assessed with the ADMET structure activity (admetSAR) algorithm [52] [Table S3]. Despite good docking results, lanreotide, mitoxantrone, octreotide, pasireotide, teniposide and midostaurin have severe side effects, suggesting that they should not be used under normal circumstances. Of the highest ranked drugs across the case studies, only midostaurin was found to have physiochemical properties that would limit effectiveness in humans.

We note that lanreotide, octreotide, pasireotide and somatostatin were excluded from SWISSDOCK analysis as they were irretrievable from its associated database. Moreover, these drugs were not in clinical trials and other than somatostatin all these other drugs had considerable side effects as assessed by the ADMET analysis.

Discussion

Our study shows that by applying our pipeline with its array of variety of multivariate approaches from a single-omics and multi-omics angle, we were able to identify a list of 84 potential drug–target combinations for COVID-19 across two separate case studies on two different systems (Figure 7, Figure S17). In case study 1, our findings suggested that a small but distinct set of proteins was sufficient to represent the cell state 24 hours post-viral infection and that late-stage viral infection was the main source of variation in the data. Similarly, in case study 2, a subset of multi-omics features characterised the state of the patient (less severe, more severe). Both sets of features overlapped with highly scoring features identified in both independent analyses. Supplementing these findings with an ensemble of *in silico* molecular docking methods and pharmacokinetic screens further narrowed down this list to seven drugs in total across both case studies: amsacrine, bosutinib, ceritinib, crizotinib, nintedanib, midostaurin and sunitinib target and inhibit various kinases.

However, in the original publication from which the source data for case study 1 was obtained, pladienolide B, 2-deoxy-d-Glucose, ribavirin and NMS-873 were identified as potential drugs of interest [3], which target various components of the RNA processing and protein machinery, along with glycolysis. In contrast, our list of highest scoring candidates is the anti-cancer drugs amsacrine, ceritinib and crizotinib targeting cell cycle and cell growth pathways, which is supported by some experimental [20] and theoretical evidence [5, 34]. In the first scenario, inhibiting cell metabolism, RNA splicing and translation would limit many biological processes in the cell, including that of some viruses [15]. In the second scenario, cell growth and cell cycle pathways are inhibited, which are known to limit replication of some viruses such as Ebola and influenza [21, 28], including SARS-CoV-2 [20]. While different results are obtained, in each scenario, experimental evidence supports these and further illustrates how multi-omics data harmonisation views data at an angle that is not normally accessible. The authors of case study 2 did not directly conduct drug discovery and therefore provided no screened drugs for us to compare ours to.

It is important to note that an *in vitro* human cancer cell line was used in case study 1. Cancer cell lines, as well as *in vitro* cell cultures, experience distinct physiological and environmental conditions compared to cells in a human body. Furthermore, the possibility that these anti-cancer drugs were identified due to the nature of system under study cannot be ruled out, though cell growth pathways are indeed shown to be upregulated in SARS-CoV-2 infection [20]. We also analysed an additional case study on a system not involving cancer cells but patients [31], with similarities in results.

To place our results in a broader context, we also compared our analysis of case study 1 [3] with two other multi-omics studies that identified SARS-CoV-2 phosphorylation sites [11] and SARS-CoV-2 RNA modification sites [19]. Although direct comparison of our results with these published data is not possible, we note that all these studies highlight several features of interest. ORF1a polyprotein, S glycoprotein, Accessory protein

3a, Nucleocapsid protein, non-structural protein 3 and non-structural protein 12 as well as their associated genes are implicated to have a degree of regulation, suggesting that these have important roles in viral biology. Therefore, this further illustrates that multi-omics perspectives reveal previously unknown RNA and protein modifications, which most studies do not account for. Furthermore, investigating this less-studied regulatory layer may yield further information critical to our understanding of SARS-CoV-2 biology and drug targeting.

Unwanted variation from technical sources or from repeated measurements on the same individual is common in biological datasets, which may decrease the accuracy of results or lead to false conclusions [25]. In case study 1, the main biological sources of variation in the data were masked, requiring a modified approach. Hence, accounting for repeated measurements on the same samples [26], or batch effects (as is the case in many publicly available datasets) caused by experimental design, is necessary before downstream data analysis. Individual omics datasets intended for any analyses should be first assessed individually to avoid misleading conclusions.

Conclusion

We demonstrate with two independent case studies that our multi-modal data harmonisation pipeline easily generates a list of pertinent biological features for downstream analyses without using algorithms specifically tailored to the datasets. With its wide scope for input data independent of the system under study, we applied it to accelerate our understanding of a medically relevant biological system. Our method supplements the valuable information obtainable with single-omics data and creates new avenues for screening potential treatments due to its cross-omics system-level perspective. Furthermore, in cases where matched samples across omics data of two or more modalities are available, our pipeline can also harmonise future datasets regardless of the system under study. We emphasise that it does not require customisation to fit a specific problem other than parameter tuning.

Key Points

- First, we demonstrate with COVID-19 case studies that to develop biomarkers from a system-level perspective, a holistic cross-omics approach is needed for studying complex phenotype.
- Secondly, we present an open-access, reproducible, flexible, machine learning multivariate approach to extract relevant information from noisy and heterogeneous multi-omics data.
- To show the effectiveness and scope of our pipeline, we apply it to high-throughput multi-omics data on SARS-CoV-2, recover published results and identify potential therapeutic drugs as well as targets.

Supplementary Data

Supplementary data are available online at *Briefings in Bioinformatics*.

Availability of source code and requirements

All documentation, steps, code, parameters, command line arguments and software versions used to generate results in this paper are publicly available in an open source software repository hosted on gitlab at https://gitlab.com/tyagilab/sars-cov-2 under a MIT license. Documentation is available at the same location, licensed as CC-BY-3.0 AU. The snapshot of the repository that produced the data generated in this paper is available at a stable DOI [8].

Availability of supporting data and materials

Primary data were generated by third parties and are publicly available [3, 31]. For case study 1, translatome data are available from the source publication as SupplementaryTable1 and proteome data are available as SupplementaryTable2. For case study 2, the authors provided their data in ansqldatabase.

Ethical approval

Not applicable.

Competing interests

There is no competing interest.

Funding

Australia–India Strategic Research Fund (AISRF) Early- and Mid-Career Researcher (EMCR) Fellowship by the Australian Academy of Science and Australian Women Research Success Grant at the Monash University (to S.T.); the Australian Government Research Training Program Scholarship and the Monash Faculty of Science Dean's Postgraduate Research Scholarship to T.C.; the National Health and Medical Research Council Career Development fellowship (GNT1159458 to K.-A.L.C.).

Author contributions statement

Conceptualisation—S.T., T.C.; formal analysis—K.-A.L.C., M.P., S.T., T.C.; funding acquisition—S.T.; investigation—K.-A.L.C., M.P., S.T., T.C.; resources—S.T.; supervision—K.-A.L.C., S.T.; validation—S.T., M.P., T.C.; visualisation—M.P., S.T., T.C.; writing (original draft)—M.P., S.T., T.C.; writing (review and editing)—K.-A.L.C., M.P., S.T., T.C.

Acknowledgments

We thank DavidA.Matthews for helpful discussions and feedback. We thank YashpalRamakrishnaiah for performing an extended analysis of the primary data used in this article. The authors thank the HPC team at the Monash eResearch Centre for their continuous personnel support. This work was supported by the MASSIVEHPCfacility. WeacknowledgeandpayrespectstotheEldersandTraditionalOwnersofthelandonwhichourfourAustraliancampusesstand.

References

1. Argelaguet R, Velten B, Arnol D, *et al*. Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol* 2018; **14**(6): 1–13.
2. Benevento M, Tonge PD, Puri MC, *et al*. Proteome adaptation in cell reprogramming proceeds via distinct transcriptional networks. *Nat Commun* 2014; **5**.
3. Bojkova D, Klann K, Koch B, *et al*. Proteomics of SARS-CoV-2-infected host cells reveals therapy targets. *Nature* 2020; **583**: 469–472.
4. Burley SK, Berman HM, Christie C, *et al*. RCSB Protein Data Bank: sustaining a living digital data resource that enables breakthroughs in scientific research and biomedical education. *Protein Sci* 2018; **27**(1): 316–30.
5. Cava C, Bertoli G, Castiglioni I. A protein interaction map identifies existing drugs targeting SARS-CoV-2. *Res Square SARS-Cov-2 Preprints* 2020; **21**(1).
6. Chen L, Xu J, Li SC. DeepMF: Deciphering the latent patterns in omics profiles with a deep learning method. *BMC Bioinform* 2019; **20**(Suppl 23): 1–13.
7. Chen Q, Allot A, Lu Z. Keep up with the latest coronavirus research. *Nature* 2020; **579**(7798): 193.
8. Chen T, Philip M, Lê Cao K-A, Tyagi S. Multi-omics data harmonisation for the discovery of COVID-19 drug targets. 2021. https://doi.org/10.5281/zenodo.4562010.
9. Chen T, Tyagi S. Integrative computational epigenomics to build data-driven gene regulation hypotheses. *GigaScience* 2020; **9**(6): 1–13.
10. Chen Y-M, Zheng Y, Yu Y, *et al*. COVID-19 severity is associated with immunopathology and multi-organ damage. *medRxiv* 2020;2020.06.19.20134379.
11. Davidson AD, Williamson MK, Lewis S, *et al*. Characterisation of the transcriptome and proteome of SARS-CoV-2 using direct RNA sequencing and tandem mass spectrometry reveals evidence for a cell passage induced in-frame deletion in the spike glycoprotein that removes the furin-like cleavage site. *bioRxiv* 2020;2020.03.22.002204.
12. Dimitrakopoulos C, Hindupur SK, Hafliger L, *et al*. Network-based integration of multi-omics data for prioritizing cancer genes. *Bioinformatics* 2018; **34**(14): 2441–8.
13. Forli S, Huey R, Pique ME, *et al*. Computational protein–ligand docking and virtual drug screening with the AutoDock suite. *Nat Protoc* 2016; **11**(5): 905–19, 5.
14. Grosdidier A, Zoete V, Michielin O. SwissDock, a protein-small molecule docking web service based on EADock DSS. *Nucleic Acids Res* 2011; **39**(SUPPL. 2): 270–7.
15. Gualdoni GA, Mayer KA, Kapsch AM, *et al*. Rhinovirus induces an anabolic reprogramming in host cell metabolism essential for viral replication. *Proc Natl Acad Sci U S A* 2018; **115**(30): E7158–65.
16. Heyes MP, Saito K, Crowley JS, *et al*. Quinolinic acid and kynurenine pathway metabolism in inflammatory and non-inflammatory neurological disease. *Brain* 1992; **115**(5): 1249–73.
17. Hoffmann M, Kleine-Weber H, Krueger N, *et al*. The novel coronavirus 2019 (2019-nCoV) uses the SARS-coronavirus receptor ACE2 and the cellular protease TMPRSS2 for entry into target cells. *bioRxiv* 2020;2020.01.31.929042.
18. Hussein SMI, Puri MC, Tonge PD, *et al*. Genome-wide characterization of the routes to pluripotency. *Nature* 2014; **516**(7530): 198–206.
19. Kim D, Lee JY, Yang JS, *et al*. The architecture of SARS-CoV-2 transcriptome. *Cell* 2020; **181**(4): 914–921.
20. Klann K, Bojkova D, Tascher G, *et al*. Growth factor receptor signaling inhibition prevents SARS-CoV-2 replication. *bioRxiv* 2020;2020.05.14.095661.
21. Kumar N, Liang Y, Parslow TG, *et al*. Receptor tyrosine kinase inhibitors block multiple steps of influenza A virus replication. *J Virol* 2011; **85**(6): 2818–27.
22. Labbé CM, Rey J, Lagorce D, *et al*. MTiOpenScreen: a web server for structure-based virtual screening. *Nucleic Acids Res* 2015; **43**(W1): W448–54.
23. Lê Cao KA, Boitard S, Besse P. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinform* 2011; **12**.
24. Lê Cao KA, Rossouw D, Robert-Granié C, Besse P. A sparse PLS for variable selection when integrating omics data. *Stat Appl Genet Mol Biol* 2008; **7**(1).
25. Leek JT, Scharpf RB, Bravo HC, *et al*. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet* 2010; **11**(10): 733–9.
26. Liquet B, Lê Cao KA, Hocini H, Thiébaut R. A novel approach for biomarker selection and the integration of repeated measures experiments from two assays. *BMC Bioinform* 2012; **13**(1): 1–14.
27. Lock EF, Hoadley KA, Marron JS, *et al*. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Ann Appl Stat* 2013; **7**(1): 523–42.
28. Luthra P, Aguirre S, Yen BC, *et al*. Topoisomerase II inhibitors induce DNA damage-dependent interferon responses circumventing Ebola virus immune evasion. *mBio* 2017; **8**(2).
29. Migaud M, Gandotra S, Chand HS, *et al*. Metabolomics to predict antiviral drug efficacy in Covid-19. *Am J Resp Cell Mol Biol* 2020; **63**(3): 396–8.
30. Mossel EC, Huang C, Narayanan K, *et al*. Exogenous ACE2 expression allows refractory cell lines to support severe acute respiratory syndrome coronavirus replication. *J Virol* 2005; **79**(6): 3846–50.
31. Overmyer KA, Shishkova E, Miller IJ, *et al*. Large-scale multi-omic analysis of COVID-19 severity. *Cell Syst* 2021; **12**(1): 23–40.e7.
32. The pandas development team. *pandas-dev/pandas: Pandas*. 2020.
33. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2020.
34. Rajagopal K, Varakumar P, Aparna B, *et al*. Identification of some novel oxazine substituted 9-anilinoacridines as SARS-CoV-2 inhibitors for COVID-19 by molecular docking, free energy calculation and molecular dynamics studies. *J Biomol Struct Dyn* 2020; **0**(0): 1–12.
35. Rohart F, Gautier B, Singh A, *et al*. mixOmics: an R Package for omics feature selection and multiple data integration. *PLoS Comput Biol* 2017; **13**(11): e1005752.
36. Sánchez-Linares I, Pérez-Sánchez H, Cecilia JM, *et al*. High-throughput parallel blind virtual screening using BINDSURF. *BMC Bioinform* 2012; **13**(SUPPL 1).
37. Schneidman-Duhovny D, Inbar Y, Nussinov R, *et al*. PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Res* 2005; **33**(SUPPL. 2): 363–7.
38. Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* 2009; **25**(22): 2906–12.
39. Singh A, Shannon CP, Gautier B, *et al*. DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics* 2019; **35**(17): 3055–62.

40. Sofia MA, Ciorba MA, Meckel K, *et al*. Tryptophan metabolism through the kynurenine pathway is associated with endoscopic inflammation in ulcerative colitis. *Inflamm Bowel Dis* 2018; **24**(7): 1471–80.
41. Stuart T, Butler A, Hoffman P, *et al*. Comprehensive integration of single-cell data. *Cell* 2019; **177**(7): 1888–902.e21.
42. Tenenhaus M. *La Regression PLS: Theorie et Pratique*. Paris: Editions Technic, 1998.
43. Tonge PD, Corso AJ, Monetti C, *et al*. Divergent reprogramming routes lead to alternative stem-cell states. *Nature* 2014; **516**(7530): 192–7.
44. Ursu O, Holmes J, Bologa CG, *et al*. DrugCentral 2018: an update. *Nucleic Acids Res* 2019; **47**(D1): D963–70.
45. Wahedi HM, Ahmad S, Abbasi SW. Stilbene-based natural compounds as promising drug candidates against COVID-19. *J Biomol Struct Dyn* 2020; **0**(0): 1–10.
46. Wang B, Mezlini AM, Demir F, *et al*. Similarity network fusion for aggregating data types on a genomic scale. *Nat Method* 2014; **11**(3): 333–7.
47. Waterhouse A, Bertoni M, Bienert S, *et al*. Homology modelling of protein structures and complexes. *Nucleic Acids Res* 2018; **46**(W1): W296–303.
48. Welch JD, Kozareva V, Ferreira A, *et al*. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell* 2019; **177**(7): 1873–87.e17.
49. McKinney W. Data structures for statistical computing in Python. In: van der Walt S and Millman J (eds). *Proceedings of the 9th Python in Science Conference*, pp. 56–61, 2010.
50. Wold H. Estimation of principal components and related models by iterative least squares. In P R Kishnaiah (ed). *Multivariate Analysis*, chapter Estimation, pp. 391–420. New York: Academic Press, 1966.
51. Wold H. Path models with latent variables: the NIPALS approach. In H M Blalock (ed). *Quantitative Sociology: International Perspectives on Mathematical and Statistical Model Building*, chapter Path model, pp. 307–57. New York: Academic Press, 1975.
52. Yang H, Lou C, Sun L, *et al*. AdmetSAR 2.0: web-service for prediction and optimization of chemical ADMET properties. *Bioinformatics* 2019; **35**(6): 1067–9.
53. Yang Z, Michailidis G. A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. *Bioinformatics* 2016; **32**(1): 1–8.
54. Zhang S, Liu CC, Li W, *et al*. Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Res* 2012; **40**(19): 9379–91.